DS2000
4/11 - Tues.

Admin

- HW8 should be returned thurs
- no class on 4/18 :)
- this Fri (4/14) - wrap up day → bit.ly/ds2000-wrapup

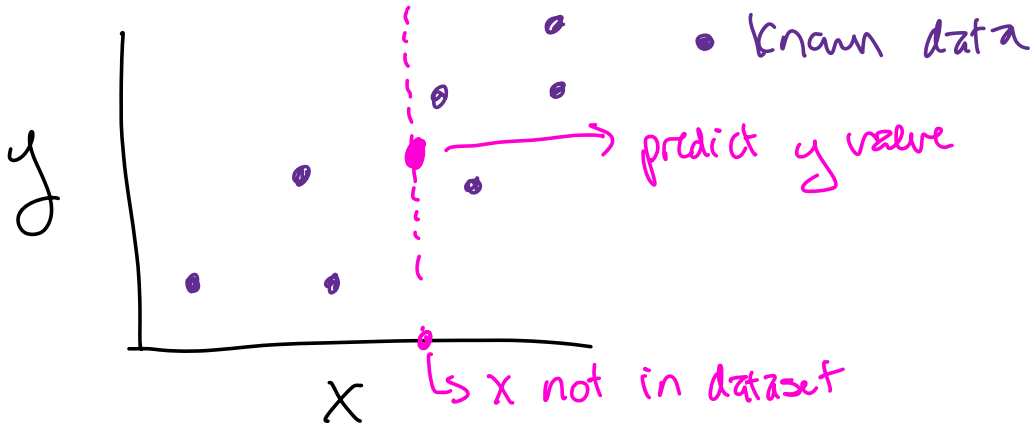
Agenda

1. Linear regression
2. Pandas, seaborn, scipy
3. Python

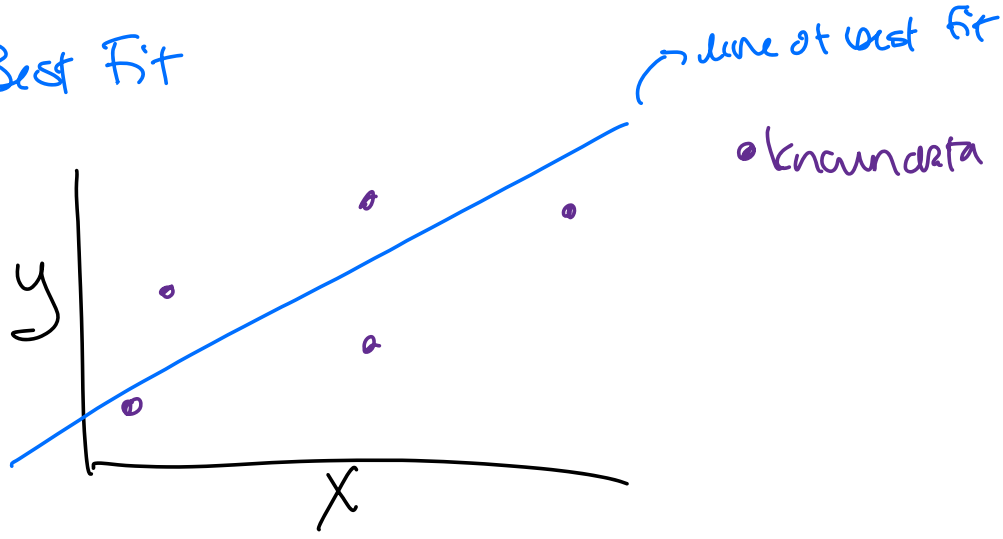
1. Linear Regression

- relationship between 2 variables
- used as prediction model
- X: independent variable
- y: dependent variable

} given x not in dataset, what would y be?



Line of Best Fit

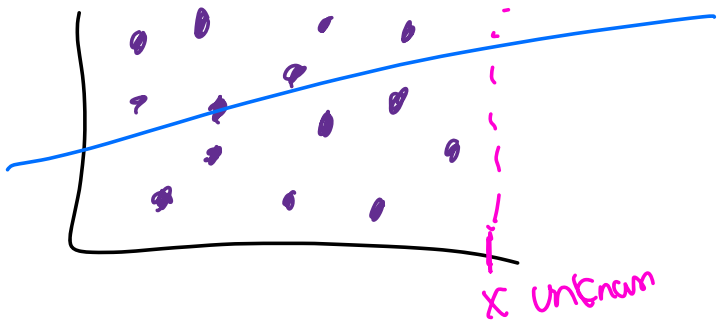


want to predict y, given x

$$y = mx + b$$

↳ slope ↳ intercept

Python will give a line of best fit for any data



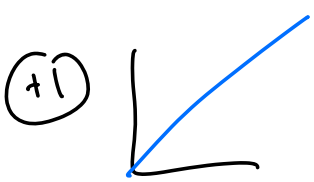
But: first step!

- What is correlation between x and y?
- If highly correlated, using the line of best fit makes sense
- default in Python: pearson

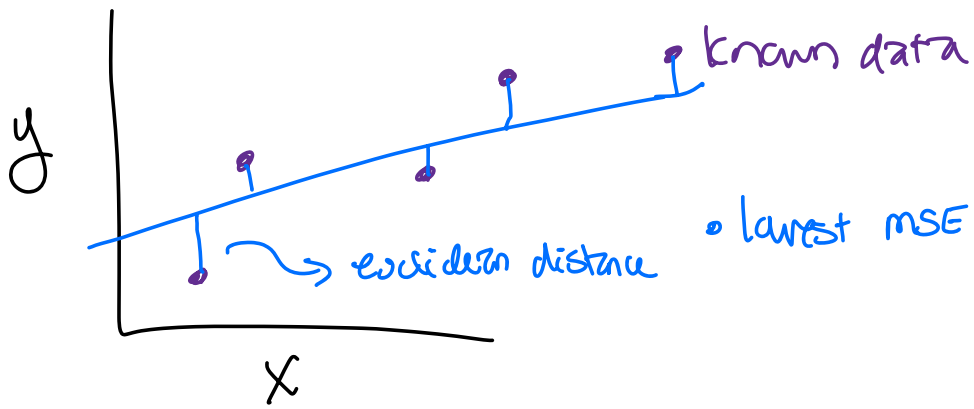
-1 to $+1$

close to zero: bad :(

close to $+1, -1$: good :)



Line of Best Fit : mean squared error (MSE)



2. Pandas / seaborn / scipy

Today's data:

- Nicolas Cage movies, precipitation in Boston
- JP property values (time series)

Python we need

Pandas — file, correlation

```
import pandas as pd

df = pd.read_csv(~)

df["Precipitation"]  ——> a column

df.corr()  ——> correlation
```

Seaborn — plot the line

```
import seaborn as sns

sns.regplot(~)  —> draws scatter + line
```

Scipy — get the slope, intercept

```
from scipy import stats

stats.linregress()  ——> get slope, int.
```